Linux Kernel

Pripremio: Dinko Korunić

Verzija: 1.0, rujan 2005.

Tijekom prezentacije

- ako što **nije jasno** - pitajte!
- ako što **nije točno** - ispravite!
- diskusija je **poželjna** i **produktivna**
- ako je **prebrzo** - tražite da se uspori!
- ako je pak **presporo** i **uspavljuje** vas - lako se ubrza sa sadržajem
- vremena je malo, sadržaja mnogo - zato su neki sadržaji samo ukratko objašnjeni

Sadržaj

- dužina trajanja: 225 minuta [5x 45 minuta]
- tip tečaja: pokazno/radni
- cjeline:
  - uvod i povijesni razvoj
  - kernel, kompilacija, instalacija, moduli, boot loaderi
  - zakrpe i dodaci, praćenje, dodavanje
  - razno :)

Uvod

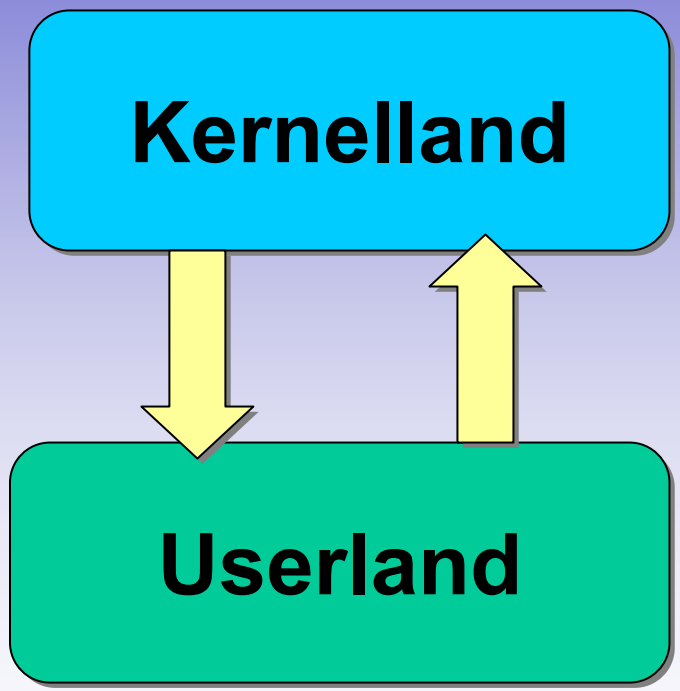
Što je kernel?

- **osnova** OS-a - ali ne i samostojeći OS!
- softver - **apstrakcija** i pristup **hardveru**, **API**, **HAL**, **upravljanje** procesima i **restrikcije** pristupa
- kategorije:
  - **monolitni**
  - **mikrokerneli**
  - **hibridi**
  - **exokerneli**

Monolitni kernel

- vrlo visoka/**bogata** apstrakcija
- **sistemski** pozivi - upravljanje procesima, konkurentnost, upravljanje memorijom
- niz **modula** - u upravljačkom načinu, vrše sistemske pozive, dijele istu memoriju, jaka integracija
- **Linux**, FreeBSD, Solaris, Windows NT
- tradicionalni Unix kerneli (BSD obitelj)

Monolitni kernel

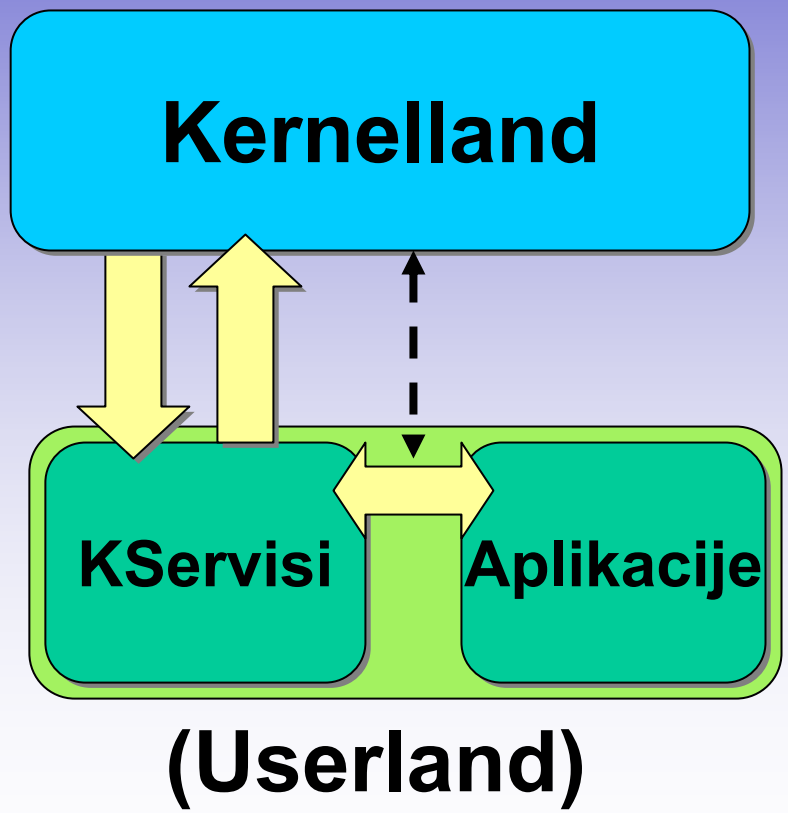


- **naizgled** najjednostavniji
- vrlo **složene** apstrakcije
- greška u jednom dijelu uzrokuje greške u **svim** dijelovima
- vrlo bliska **povezanost** koda na niskoj razini omogućava dobru **efikasnost** i **brzinu**

Mikrokerneli

- **minimalna/jednostavna** apstrakcija
- sistemski pozivi - minimalni OS **servisi** (upravljanje dretvama, adresnim prostorom i međuprocesna komunikacija)
- ostale usluge kernela - su **servisi** u **korisničkom** prostoru:
  - jednostavnije micanje nepotrebnih usluga
  - stabilnije - gasi se jedan proces
  - problem - nema garantiranog stanja!

Mikrokerneli



- lošije **performanse** od monolitnih - mnogo **kopiranja** među aplikacijama/servisima

- **context switching** je skup!

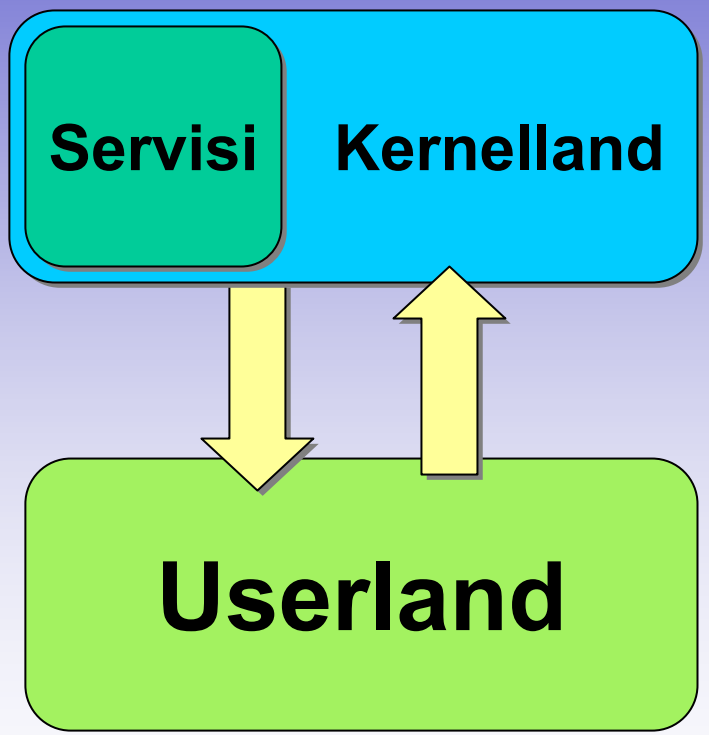
Mikrokerneli

- primjeri:
  - AIX, AmigaOS, Minix, QNX, Symbian
  - L4 obitelj
  - Mach: GNU Hurd, NextStep, OpenStep, Mac OSX
- Mikro vs. monolitno:
  - Linux vs. Tanenbaum
  - monolitni - jednostavniji **dizajn**, brži **razvoj**
  - mikrokernel - **sigurnije**, svaka od OS komponenata u vlastitom zaštićenom memorijskom prostoru

Hibridni kerneli

- varijanta mikrokernela - nešto dodatnog koda u kernelu radi brzine = **kompromis**
- Windows NT:
  - mikrokernel - kernel
  - servisi - NT executive
  - komunikacija - dijeljena memorija, LPC pozivi
- DragonFly BSD, Windows NT/2k/XP, ReactOS, BeOS

Hibridni kerneli



- što **više** koda izvan kernela

- komunikacija kroz **poruke**

- nešto **koda** unutar kernela radi brzine! (manje latencije)

Linux kernel

- **slobodan** monolitni modularni kernel nalik na Unix, alternativa Minixu
- autor **Linus Torvalds** 1991, GNU licenca
- povijest:
  - 1994 - Linux 1.0.0
  - 1996 - Linux 2.0.0
  - 1999 - Linux 2.2.0
  - 2001 - Linux 2.4.0
  - 2003 - Linux 2.6.0

Linux kernel

- verzije:
  - 3 ili 4 broja - A.B.C i opcionalno .D
  - A - osnovna verzija kernela, iznimno rijetko
  - B - glavna verzija, **parni** - stabilno i produkcijski, neparni - razvojno i nestabilno!
  - C - nekad - sigurnosne zakrpe, bugfixevi, nove mogućnosti, driveri; danas - samo kad su novi driveri ili mogućnosti (**veće** promjene!)
  - D - bugfixevi i sigurnosne zakrpe (**manje** promjene!)

Linux kernel

- ljudi - maintaineri:
  - 2.0 - David Weinehall
  - 2.2 - Alan Cox, danas Marc-Christian Petersen
  - 2.4 - Marcelo Tosatti
  - 2.6 - Linus Torvalds
  - Robert Love - preempt,VM, inotify
  - Ingo Molnar - O(1). ExecShield, RT, latencije
  - Miguel de Icaza - Gnome, Mono
  - Rik van Riel - rmap, VM

Linux kernel

- na čemu radi?
  - laptopi
  - uClinux - embedded uređaji: Palm, iPAQ, itd.
  - desktopi - nVidia, ATI, itd.
  - big iron: NUMA
  - clusteri! - OpenSSI, openMosix, itd.
  - desetine SCSI, FC i inih kontrolera
  - Infiniband!
  - desetine datotečnih sustava - ROM, RAM, diskovi, RAID, GFS, journaling, obični, itd.

Linux kernel 2.6

- konzistentnije promjene nego prije - korišten komercijalni **BitKeeper** (danas GIT)
  - preciznija kontrola, **patchetovi**, službeno stablo
  - ogroman broj promjena
- **paralelni** razvoj:
  - **arhitekture**: arm, axp, ia64, sparc, mips, ppc...
  - **forkoffovi**: mm, ac, mjb, wli, kj, aa, itd
  - **grupe** za razvoj uređaja: isdn, scsi, usb, driveri, itd.

Linux kernel 2.6

- do 2.6:
  - manjak **formalne** provjere koda
  - (The Cathedral and the Bazaar) Linus' Law: **"Given enough eyeballs, all bugs are shallow"**
  - Linux Test Project - 2000+ testova za kvalitetu, testovi **regresije**, automatizirana **noćna** testiranja
  - kasnije i Intel razvio **set testova**
- što se sve promijenilo? mnogo. 2.6 donosi najveće promjene do sada

Linux kernel 2.6

- promjene:
  - uglavnom udev umjesto devfs
  - moduli promijenjeni (.ko), novi način učitavanja
  - moćniji build sustav (brže, ispravno prepoznaje što je mijenjano već kompiliranom stablu)
  - ubrzan IO podsustav - moguće mijenjati schedulere (deadline, anticipatory, cfg, noop)
  - 64bit veličine datoteka (...) - 16TB fs-ovi
  - POSIX ACL-ovi i dodatni atributi (setfacl, getfacl..)
  - preemption!

Linux kernel 2.6

- O(1) process scheduler - daleko bolje skaliranje kad je mnogo procesa i visoko opterećenje
- HT-aware scheduler
- RML (Robert Love) dodaci za binding procesa po CPU (affinity) kako ručno tak i iz aplikacije
- prioritet (nice) procesa sada znači više nego prije - izbjegavati negativne brojeve!
- userspace i kernelspace IRQ balancer
- Futexes (Fast Userspace Mutexes)
- epoll - odgovor na kqueue i RT SIGIO

Linux kernel 2.6

- mnogo threading promjena, NPTL (Native Posix Threading Library), brzina...
- corefiles - /proc/sys/kernel/core_pattern
- OSS je izbačen, koristi se ALSA (Advanced Linux Sound Architecture), brdo novih uređaja
- AGP 3.0 podrška, HotPlug PCI, ISAPnP BIOS
- framebuffer
- IDE - izbjegavati SCSI emulaciju; uveden TCQ, SATA podržan kao SCSI
- SCSI - novih SCSI kontrolaca, MegaRAID2, itd.

Linux kernel 2.6

- 32bitni UID i PID-ovi
- USB kontroleri: UHCI, OHCI i EHCI; HID podrška
- ext3 - indeksiranje (tune2fs -O has_index i fsck po tome)
- NFS poboljšan + NFSv4
- sysfs - dodatne informacije umjesto procfs
- dodan JFS i XFS(!!) te CIFS i HugeTLB
- kernel AIO
- IPMI (Intelligent Platform Management Interface)
- detekcija CPU, microcode updates

Linux kernel 2.6

- power management - ACPI, Swusp2, CPUFreq, češći MCE
- novi LVM2
- SELinux umjesto npr. Grsecurity :(
- bridge u kernelu, ebtables
- IPsec! puni IPv6...
- NAPI za e100, tulip, e1000, itd.
- cryptoAPI
- NUMA, PAE, 64GB RAM

Linux kernel 2.6

- LKML:
  - http://www.ussg.iu.edu/hypermail/linux/kernel/index.html
- Bug Tracker:
  - http://bugzilla.kernel.org/
- Linux HQ:
  - http://www.linuxhq.com/index.html

Boot loader

- program:
  - prvi koji se izvršava nakon BIOS-a (ROM, Flash)
  - obavlja određene akcije - računalo bez OS-a ne može samo učitati OS
- boot loader, bootstrap loader
  - dovoljno funkcionalnosti da pokrene OS
  - obično i više stupnjeva: 1, 1.5, 2 - jedni druge učitavaju i pozivaju do podizanja OS-a
  - obično se učita na predefiniranu adresu
  - x86 - max 512 bajtova (446), na kraju AA55h

Boot loader

- **second** stage boot loader:
  - PC: NTLDR, LILO, **GRUB**, SmartBoot, IBM MBR
  - uglavnom veći od 512, više podrške (mreža, ATA, ATAPI, itd)
- alternative:
  - Aboot (Alpha), MILO (Alpha), SILO (Sparc), itd.
- **procedura** podizanja sustava:
  - CPU izvršava instr. sa FFFF0h adrese BIOS-a
  - POST, testiranje uređaja, bootable uređaji

Boot loader

- učitava se i izvršava boot sektor
- disk - MBR, koja je particija u tablici aktivna/sistemska
- svaka particija ima svoj vlastiti BR (boot sektor)

- **GRUB** - http://www.gnu.org/software/grub/:
  - multiboot - više različitih OS-ova (BSD, Hurd, Windows, Linux, chain loading, mrežni boot)
  - superiorna podrška - omogućava pregled datotečnih sustava!

Boot loader

- podržani: ext2/3, ReiserFS, XFS, UFS/FFS, VFAT, NTFS, JFS, TFTP, disketa
- konfiguracija: dinamička, menu.lst, izbornik...
- boot: Stage 1 (MBR), Stage 1.5 (30KB odmah iza MBR - obično prazno do 1MB) ili Stage 2 (bilo gdje)

- **LILO**:
  - standardni loader već niz godina
  - konfiguracija statički, upisuje se u stage 2

Boot loader

- nema podršku za specifične sustave - hardkodirane pozicije (ručno pokretanje LILO-a)
- stara računala - CHS, do 1023 cilindra
- nova - 32 i 48bit LBA adresiranje

- **SYSLINUX** - http://syslinux.zytor.com
  - za specifične primjene
  - SYSLINUX - FAT, floppy
  - ISOLINUX - ISO9660, CDROM

Boot loader

- **PXELINUX** - Pre-eXecution Environment
  - DHCP/BOOTP za adrese
  - nužan ROM ili dodatna disketa
  - TFTP za boot
- EXTLINUX - ext2/3
- MEMDISK - emulacija diska (0x80) u memoriji
- instalacijski mediji obično
- **ElTorito** standard:
  - emulacija diskete - SYSLINUX
  - bez emulacije - ISOLINUX

Kompilacija i instalacija

Zašto mijenjati?

- optimiranje:
  - specifično hardveru, brzina!
  - za pojedine potrebe - server, desktop, igre, itd.
  - standardni kerneli - generički, sporiji
- nadogradnja:
  - redovno - riješeni razni problemi
  - dodatne mogućnosti (performanse, sigurnost, podešenja, specifična podrška)
- radoznalost:
  - Bleeding edge, man!

Nužni preduvjeti

- hardver:
  - 2.4 kernel - 8MB RAM, 64 praktično
  - nužno 80ak MB prostora - izvorni kod
  - 400MB privremenih datoteka!
- softver:
  - GCC 2.95, 3.3... GNU make
  - binutils 2.12
  - module-init-tools (2.6.x kerneli!)
  - i još ponešto

Prije kompiliranja

- nužni paketi (apt-get install...):
  - gcc, cpp, make, binutils
  - libncurses5-dev
  - modutils, module-init-tools, procps, mount, sysvinit
  - lilo, grub
  - e2fsprogs, xfsprogs, reiserfsprogs, reiser4progs, jfsutils
- izbjegavati:
  - gcc4.0 - gcc3.3 za sada dovoljno stabilan!

Prije kompiliranja

- softver:
  - procps - PROC
  - jfsutils - JFS
  - e2fsprogs - EXT2/3
  - reiserfsprogs - Reiserfs v3
  - xfsprogs - XFS
- hardver:
  - 2GB RAM, SMP (2CPU ili barem HT)
  - e1000 kartice, SMP ploče imaju IO-APIC

Prije kompiliranja

- dodatni softver:
  - isdn4k-utils
  - nfs-utils
  - pcmcia-cs
  - ppp
  - quota-tools
  - alsa...
  - itd.

Saznavanje hardvera

- naredbe:
  - pnpdump - ISA
  - lspci - PCI
  - dodatno: lshw, /proc/cpuinfo
- nužno uvijek pripremiti listu hardvera
- najvažnije:
  - disk kontroler
  - datotečni sustav
  - tip i broj procesora

Kako do kernela

- iz **vlastite** distribucije:
  - kernel-cn
  - kernel-image
  - itd.
  - obično ima ili nema razne dodatke! generički!
- **čisti/vanilla** kernel:
  - http://www.kernel.org/
  - sekundarna mjesta: ftp.de.kernel.org, ftp.hr.kernel.org, itd

Životna odluka

- 2.4 kernel - /pub/linux/kernel/v2.4:
  - prokušan, malo **sigurnosnih** problema
  - **stari** kod - nema više aktivnog razvoja
  - **suboptimalan** za novi hardver
  - nema niza dodataka (ebtables, L7, SELinux, itd)
- 2.6 kernel - /pub/linux/kernel/v2.6:
  - niz **sigurnosnih** problema
  - potencijalni **bugovi**
  - aktivni i brz **razvoj**, hrpa **dodataka**!

Datoteke na kernel.org

- ChangeLog
  - razlike između verzija
  - obično korisno
- format datoteka:
  - patch-VERZIJA.gz - inkrementalna nadogradnja
  - linux-VERZIJA.tar.gz - cijeli kod, tar + gz
  - linux-VERZIJA.tar.bz2 - cijeli kod, tar + bzip2
  - *.sign - PGP potpisi!

Nabavka izvornog koda

- cd /usr/src
- wget http://www.kernel.org/pub/linux/kernel/v2.6/linux-2.6.13.3.tar.bz2
- tar xjf linux-2.6.13.3.tar.bz2
- cd linux-2.6.13.3

- slijedi konfiguriranje i kompiliranje!

Konfiguriranje

- apt-get install libncurses5-dev
- cd /usr/src/linux-VERZIJA
- načini konfiguriranja:
  - make config - ružno, tekstualno, pitanja
  - make menuconfig - polugrafički, preporučeno
  - make oldconfig - iz stare .config datoteke, radi i 2.4 u 2.6 i sl.
  - make xconfig - kroz X11
- sačuvajte konfiguraciju: cp .config ..

Moduli ili ne

- dijelovi kernela (nekad .o, danas .ko):
  - moraju biti u kernelu
    - osnovni dijelovi
    - povećavaju veličinu kernela
    - preveliki kernel - nemoguće bootati
  - mogu biti kao moduli
    - naredbe modprobe, lsmod, depmod
    - datoteka /etc/modules.conf
    - direktorij /etc/modutils i naredba update-modules

Moduli ili ne

    - datoteka /etc/modules - lista za učitavanje
  - mogu biti u kernelu (statički)
    - minimalno ubrzanje, danas više nema razlika

- preporuka:
  - apsolutno nužne stvari u kernel (SCSI/IDE ctrl, datotečni sustav)
  - ostalo kao moduli
  - Adore - potencijalni sigurnosni problem?

Uređivanje modula

- /etc/modules:
  - učitanje modula pri podizanju sustava:
  - statička lista
  - koriste init skripte (modutils, module-init-tools)
- /etc/modules.conf i /etc/modutils/*
  - definiranje odnosa modul-uređaj, akcija pri učitavanju i sl.
  - update-modules - generira modules.conf iz direktorija
  - alias, post-install, post-remove, options, itd

Opcije kernela

- nužno pregledati i proći par puta
- osnovni odjeljci:
  - code maturity, loadable modules, processor type, general setup
- uređaji:
  - ata/ide..., multi-device, block devices, scsi devices
- mreža:
  - networking
- etc.

Kompiliranje

- 2.4:
  - make dep bzImage modules modules_install
  - cp System.map /boot/System.map-verzija
  - cp arch/i386/boot/bzImage /boot/vmlinuz-verzija
- 2.6:
  - make bzImage modules modules_install
  - cp System.map /boot/System.map-verzija
  - cp arch/i386/boot/bzImage /boot/vmlinuz-verzija

LILO

- /etc/lilo.conf
- uvijek nužno ručno pokrenuti nakon promjene kernela!
- primjer (RAID):
  - lba32
  - compact
  - boot=/dev/sda
  - root=current
  - delay=5

LILO

- timeout=150
- vga=normal
- default=Linux
- image=/boot/vmlinuz
- label=Linux
- read-only
- image=/boot/vmlinuz.old
- label=LinuxOLD
- read-only
- optional

LILO

- podržava i Linux RAID 1:
  - boot=/dev/md0
  - root=/dev/md0
  - raid-extra-boot=/dev/sda,/dev/sdb

- konvencije:
  - zgodno raditi simboličke linkove u / (relativni!)
  - kerneli i mape u /boot
  - uvijek imajte fallback! čuvajte stari kernel!

GRUB

- priprema:
  - apt-get install grub
  - mkdir /boot/grub
  - cp /lib/grub/i386-pc/* /boot/grub
- GRUB shell - naredba grub
- ručna instalacija:
  - grub
  - root (hd0,3)
  - setup (hd0)

GRUB

- automatska:
  - grub-install /dev/sda
- imenovanje:
  - diskovi počinju od 0, particije od 0
  - SCSI i IDE su uvijek hd*
  - TFTP - net
  - floppy - fd*
- konfiguracija:
  - nije nužna, inače /boot/grub/menu.lst

GRUB

- primjer:
  - default 1
  - timeout 5
  - title Dell Utility
  - root (hd0,0)
  - makeactive
  - chainloader +1
  - title Debian GNU/Linux
  - root (hd0,1)
  - kernel /boot/vmlinuz root=/dev/md0 reboot=warm

Zakrpe i dodaci

Sysctl

- sučelje prema jezgri, kao i procfs
- omogućava fino podešavanje jezgrinih parametara

- konfiguracija:
  - datoteka /etc/sysctl.conf
  - standardne vrijednosti nisu optimalne za sve sustave
  - kernel-cn donosi uglavnom vlastite postavke

Sysctl podešenja

- primjeri:
  - net.ipv4.conf.all.accept_redirects=0
  - net.ipv4.conf.all.accept_source_route=0
  - net.ipv4.conf.all.log_martians=1
  - net.ipv4.conf.all.rp_filter=1
  - net.ipv4.conf.all.secure_redirects=1
  - net.ipv4.conf.all.send_redirects=0
  - net.ipv4.icmp_echo_ignore_broadcasts=1
  - net.ipv4.icmp_ignore_bogus_error_responses=1

Sysctl podešenja

- net.ipv4.ip_forward=0
- net.ipv4.ip_local_port_range=10000 65000
- net.ipv4.tcp_ecn=0
- net.ipv4.tcp_max_syn_backlog=8192
- net.ipv4.tcp_retries1=2
- net.ipv4.tcp_rfc1337=1
- net.ipv4.tcp_syncookies=1
- vm.min_free_kbytes=10240
- dev.rtc.max-user-freq=1024
- vm.swappiness=75

Različiti dodatci

- Ketchup:
  - http://www.selenic.com/ketchup/
  - automatski patching i praćenje sourceva
  - nužan Python i GNUPG
- L7 Netfilter:
  - http://l7-filter.sourceforge.net/
  - aplikativno filtriranje, regularni izrazi
  - primjerice P2P filtriranje
  - potrebni i zakrpani Iptables

Različiti dodatci

- Grsecurity:
  - http://grsecurity.net/
  - neslužbeno: http://www.grsecurity.net/~spender/
  - PaX, ASLR, itd.
  - potreban paxctl, chpax, gradm2
- Con Colivas:
  - http://members.optusnet.com.au/ckolivas/kernel/
  - interaktivnost

Različiti dodatci

- Andrew Morton:
  - http://www.kernel.org/pub/linux/kernel/people/akpm/mm/
  - ultra-svježi dodaci
- Usermode Linux:
  - http://user-mode-linux.sourceforge.net/
  - virtualni strojevi
  - potreban i program/userland kernel

Različiti dodatci

- Software Suspend 2:
  - http://www.suspend2.net/
  - suspendiranje na disk (hibernacija)

- clusteri:
  - OpenSSI:
    - http://www.openssi.org
  - openMosix:
    - http://openmosix.sourceforge.net/

Patchiranje

- reverzibilna operacija
- postupak:
  - cd /usr/src/linux-VERZIJA
  - gzip -dc ../patch-NESTO.gz | patch -p1
  - ili
  - patch -p1 ../patch-NESTO
- čuvajte uvijek čisti kod

Grsecurity

- dodatci da bi jezgra bila sigurnija
  - stog nije izvršan! greške u programima teže iskoristive
  - ACL radi kontrole po procesima
  - ASLR, randomizacija PIDova, forkbomb zaštite
  - randomizacija IP IDjeva, TTLova, TCP izvornih portova
  - coredump shema
  - sysctl za upravljanje
  - itd

Layer 7 Netfilter

- idealno za routere/gatewaye
- filtriranje ili markiranje prometa
- pravila - regularni izrazi, nadograđuju se
- primjer:
  - -A FORWARD -m layer7 --l7proto gnutella -j REJECT --reject-with icmp-port-unreachable
  - -A FORWARD -m layer7 --l7proto bittorrent -j REJECT --reject-with icmp-port-unreachable

Kernel-package

- apt-get install kernel-package
- primjer:
  - make-kpkg kernel_image
  - make-kpkg --rootcmd fakeroot kernel_image

- priprema skripte, loader, gradi Debian paket

Diskusija!